Published online 21 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D211–D215***
doi:10.1093/nar/gkn785

InterPro: the integrative protein signature database

Sarah Hunter[1,*], Rolf Apweiler[1], Teresa K. Attwood[2], Amos Bairoch[3], Alex Bateman[4], David Binns[1], Peer Bork[5], Ujjwal Das[1], Louise Daugherty[1], Lauranne Duquenne[6], Robert D. Finn[4], Julian Gough[7], Daniel Haft[8], Nicolas Hulo[3], Daniel Kahn[6], Elizabeth Kelly[9], Aurélie Laugraud[6], Ivica Letunic[5], David Lonsdale[1], Rodrigo Lopez[1], Martin Madera[7], John Maslen[1], Craig McAnulla[1], Jennifer McDowall[1], Jaina Mistry[4], Alex Mitchell[1,2], Nicola Mulder[9], Darren Natale[10], Christine Orengo[11], Antony F. Quinn[1], Jeremy D. Selengut[8], Christian J. A. Sigrist[3], Manjula Thimma[1], Paul D. Thomas[12], Franck Valentin[1], Derek Wilson[13], Cathy H. Wu[10] and Corin Yeats[11]

[1]EMBL Outstation European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, [2]Faculty of Life Science and School of Computer Science, The University of Manchester, Manchester, UK, [3]Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland, [4]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [5]European Molecular Laboratory (EMBL), Heidelberg, Germany, [6]Pole Rhone-Alpins de BioInformatique (PRABI) and Laboratoire de Biométrie et Biologie Evolutive, CNRS, INRIA, Université de Lyon, Université Lyon 1, Villeurbanne, France, [7]Department of Computer Science, University of Bristol, Woodland Road, Bristol, UK, [8]J. Craig Venter Institute (JCVI), Rockville, MD, 20850, USA, [9]Computational Biology Unit, University of Cape Town, South Africa, [10]Protein Information Resource (PIR), Georgetown University Medical Center, Washington, DC, USA, [11]Department of Structural and Molecular Biology, University College London, London, UK, [12]Evolutionary Systems Biology, SRI International, Menlo Park, CA, 94025-3493, USA and [13]MRC Laboratory of Molecular Biology, Cambridge

Received September 15, 2008; Revised October 8, 2008; Accepted October 9, 2008

ABSTRACT

The InterPro database (http://www.ebi.ac.uk/interpro/) integrates together predictive models or 'signatures' representing protein domains, families and functional sites from multiple, diverse source databases: Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY and TIGRFAMs. Integration is performed manually and approximately half of the total ~58 000 signatures available in the source databases belong to an InterPro entry. Recently, we have started to also display the remaining un-integrated signatures via our web interface. Other developments include the provision of non-signature data, such as structural data, in new XML files on our FTP site, as well as the inclusion of matchless UniProtKB proteins in the existing match XML files. The web interface has been extended and now links out to the ADAN predicted protein–protein interaction database and the SPICE and Dasty viewers. The latest public release (v18.0) covers 79.8% of UniProtKB (v14.1) and consists of 16 549 entries. InterPro data may be accessed either via the web address above, via web services, by downloading files by anonymous FTP or by using the InterProScan search software (http://www.ebi.ac.uk/Tools/InterProScan/).

INTRODUCTION

InterPro (1) is an integrative database which was founded 10 years ago when the PROSITE (2), PRINTS (3), Pfam (4) and ProDom (5) databases formed a consortium to amalgamate the predictive signatures they individually produced into a single resource. Since then, six other member databases have also joined and their data has been integrated: SMART (6), TIGRFAMs (7), PIRSF (8), SUPERFAMILY (9), PANTHER (10) and Gene3D (11).

*To whom correspondence should be addressed. Tel: +44 0 1223 494 481; Fax: +44 0 1223 494 468; Email: hunter@ebi.ac.uk

The signatures of each member database are built using different but complementary methodologies.

When different signatures match the same set of proteins in the same region on the sequence, they are presumed to be describing the same functional family, domain or site and are placed into a single InterPro entry by a curator. Grouping equivalent signatures from different sources together in this way has obvious benefits, giving signatures consistent names and annotation. It also highlights potentially erroneous signature hits. One would expect that remote homologues might only match a single signature from a multiple signature entry but these outliers could also be explained by single matches being false positive, hence the user should regard these results more cautiously.

Collectively considering the total set of signatures from the member databases also increases overall coverage of protein space. The coverage of various sequence databases by InterPro signatures is shown in Table 1. InterPro signature matches to the UniProt Knowledgebase [UniProtKB; (12)] are regularly calculated using the InterProScan software package (13) and this information is used to aid UniProtKB curators in their annotation of Swiss-Prot proteins, as well as being the basis of the automatic systems which add annotation to UniProtKB/TrEMBL (12). The UniParc protein archive and UniMES meta-genomic sequence databases (14) are also put through InterPro analysis pipelines and many genomic sequencing projects continue to use InterPro and its software to functionally characterize whole genomes (15,16).

If a signature only matches a subset of proteins compared to another signature, it is likely that this signature is more functionally or taxonomically specific than the other. In this case, the signatures would be deemed to be related; the signature matching the subset would be termed a child, the other signature being its parent. These parent–child relationships are created by InterPro's curators during the integration process and a hierarchy of how the integrated signatures relate to each other is thus constructed. In this way, InterPro also increases the depth of annotation of protein space.

Once an InterPro entry is created, curators add annotation, such as a descriptive abstract, name and cross-references to other resources, including Gene Ontology (GO) terms (17). Semi-automatic procedures create and maintain links to an array of other databases, including the protease resource MEROPS (18), the protein interaction database IntAct (19), the protein sequence clusters in CluSTr (20) and the 3D protein structure database PDB (21). Additionally, if a protein has a solved 3D structure in PDB or a structure modelled in either the MODBASE (22) or SWISS-MODEL (23) databases, this information is shown together with the member databases' signature matches in the graphical display on the InterPro Web interface.

Users are able to access all pre-computed matches of signatures to UniProtKB via the web interface in a variety of graphical and text-based formats. They can change how these matches are shown by either sorting by UniProtKB identifier or name, for example, or by electing to display matches based on their taxonomy, solved 3D structures or splice variants. They can also download XML-format files of matches to UniProtKB, the UniProt Archive (UniParc) and UniMES meta-genomic sequence database.

InterProScan is made available via the web at http://www.ebi.ac.uk/Tools/InterProScan/, and the entire package can be downloaded from the FTP site ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/index.html. InterProScan allows users to submit their own sequences to the search algorithms and processing from InterPro and its member databases. They can receive results in various formats showing the signatures that match their sequence(s), the InterPro entry (if any) into which each signature is integrated and any GO terms associated with those entries. SOAP-based web services also exist (http://www.ebi.ac.uk/Tools/webservices/WSInterProScan.html) which allow users to submit their own nucleotide and protein sequences programmatically (24).

NEW FEATURES IN InterPro

Annotation

InterPro curators continue to integrate new signatures from member databases into entries. The entries are classified according to the type of signature they group together. Previously, the categories comprised family, domain, repeat, post-translational modification (PTM), active site and binding site. A new type has recently been introduced called 'conserved site' which covers any PROSITE patterns which are not a PTM or do not have a binding or catalytic activity but are conserved across members of a protein family.

Protein matches and XML files

Matches of InterPro signatures to UniProtKB, UniParc and UniMES databases are continuously calculated. Each unique protein sequence is stored only once in

Table 1. Coverage of the major sequence databases UniProtKB, UniParc and UniMES by InterPro signatures

Sequence database	Number of proteins in database	Number of proteins with >0 matches to InterPro	Number of proteins with >0 matches combined member database signatures
UniProtKB/Swiss-Prot	397 539	369 830 (93.0%)	379 897 (95.6%)
UniProtKB/TrEMBL	6 212 793	4 628 221 (74.5%)	4 894 258 (78.8%)
UniProtKB (Total)	6 610 332	4 998 051 (75.6%)	5 274 155 (79.8%)
UniParc	17 718 252	12 211 006 (68.9%)	13 290 858 (75.0%)
UniMES	6 028 191	4 132 464 (68.6%)	4 461 935 (74.0%)

The number of proteins matching signatures from InterPro and those matching the full set of member database signatures are shown.

UniParc and so, to minimize calculation overhead, searches are run cumulatively; only once per signature per unique sequence. Consequently, we can now offer pre-computed match data for all ~17 million sequences currently in UniParc via our FTP site files. This total includes UniMES sequences, which are also provided in a separate file. Supplementary statistics about the release version of each member database and number of signatures are also now in the XML files.

A new file (feature.xml) has been created which contains non-signature match data from the structural databases (PDB, MODBASE and SWISS-MODEL) for UniProtKB proteins. Proteins from UniProtKB that do not match any of the signatures in InterPro's member databases have been added to our match XML file. Previously these were omitted to save space, however, their inclusion enables users to check whether a set of pre-computed matches for a particular protein is missing because no signatures were found to match the protein or because it has not yet been analysed by the match pipeline. All our XML and flat files are updated when InterPro is publicly released, which is currently a cycle of ~3 months.

A new version of the InterProScan software (v4.4) has recently been released which has been modified to reflect alterations in the ways that matches are calculated by the member databases, as well as improving the indexing of the match XML files for retrieving pre-calculated matches for submitted sequences. The full set of changes in version 4.4 is detailed in the InterProScan software release notes (ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/ReleaseNotes.txt).

Web interface

No new member databases have been added to InterPro since the previous publication (1), but signatures from all the existing member databases continue to be integrated into new and existing InterPro entries. However, a large proportion (>50%) remain un-integrated. Previously, information about these un-integrated signatures was only available via the FTP site in XML files but now these signatures are displayed via the web interface on individual signature pages. Signature pages contain a minimal amount of information about the member database methods, such as their name and abstract if they are available, together with a brief description of their source database and a link back to the source database's home page. The total number of UniProtKB proteins the signature matches is shown and can be displayed by following a hypertext link.

InterPro entry pages featuring curator-integrated signatures contain annotation data such as an abstract and database cross-references. These entry pages also contain a 'taxonomic wheel', which displays the number of protein sequences from major taxonomic groups which are matched by that entry. Each taxonomic group is hyperlinked, providing taxonomic and sub-classification data, a graphical display of the proteins with respect to all signature matches and the ability to download the sequences in FASTA format.

Database cross-references

A total of 386 links have been added from the protein match pages to the ADAN database (http://adan-embl.ibmc.umh.es/). ADAN contains predicted protein–protein interactions of globular domains. Links in InterPro have also been made to DAS-related tools such as the SPICE 3D structure viewer (25) and the Dasty client (26). SPICE is a Java-based DAS client which displays protein sequences as 3D structures, together with structure and function-related data from various DAS sources. Dasty is a more general DAS client which visualizes DAS annotations on the sequence as well as other, non-positional information. The approximately 27 000 citations referenced in abstracts and in the additional reading section now link to the CiteXplore literature search tool (http://www.ebi.ac.uk/citexplore/).

Web services

New SOAP-based Web Services have been added to complement the existing InterProScan Web Service. These allow users to programmatically retrieve InterPro entry data such as the abstract, integrated signature lists or GO terms. Users can download a range of clients from http://www.ebi.ac.uk/Tools/webservices/clients/dbfetch, including PERL, C# .NET and Java clients, to access this data.

AVAILABILITY

The database and related software are freely available to be downloaded and distributed, so long as the appropriate Copyright notice is supplied (as described in the accompanying Release Notes). Data can be downloaded in a flat-file format (XML), as an Oracle database dump and via the web interface and web services mentioned in the text.

DISCUSSION

In the early stages of InterPro's evolution, signature development between the member databases was not a coordinated effort and resulted in a high level of redundancy, with some InterPro entries eventually containing up to 10 signatures. Through the collaborative efforts of the InterPro consortium, however, the amount of redundancy in signatures between the member databases is decreasing, providing more unique and valuable coverage of protein sequence data. Each database is cultivating its own niche in signature development, with the aim of expanding subfamilies and building signatures representative of newly characterized families, rather than duplicating work. This trend is illustrated in Figure 1. Thus, the future focus within InterPro will be on how signatures from different databases relate to one another within biologically informative hierarchies, rather than on simply reducing redundancy.

InterPro has shown its importance as a functional classification tool, not only through its use in high-profile sequence databases and genomics projects, but also by the number of users who access the resource and its associated services via the web. In 2008, the EBI-hosted

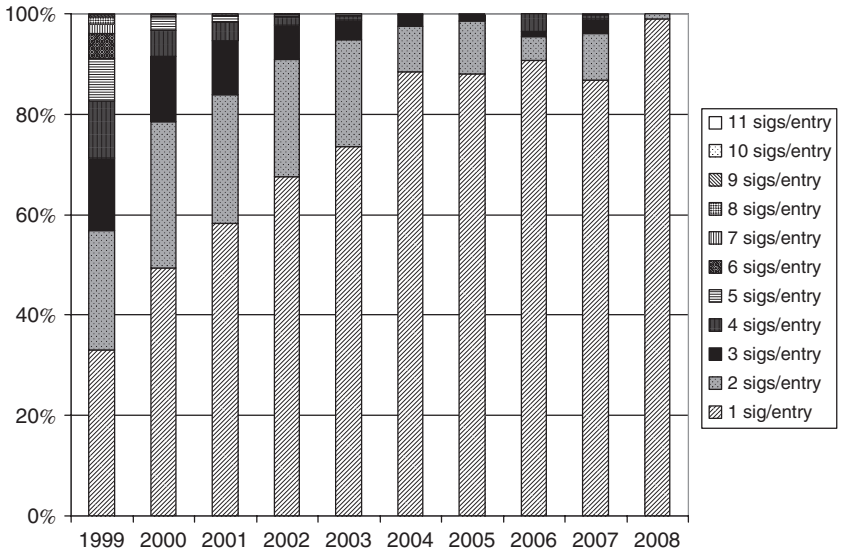

Figure 1. Trends in number of signatures integrated into a single entry, categorized by the year the entry was first created. Initially, these entries would have only contained signatures from the founding four consortium members. However, as other member databases joined, they also may have had signatures covering the same families and domains which consequently also became integrated into these entries, leading to the totals we see today. Note that the number of signatures integrated in a single year can vary (between 1000 and 5000 signatures) dependent on the member databases' release cycles.

version of InterProScan averaged over 500 000 searches a month, of which 94% were submitted via the InterProScan web service. Hundreds of copies of the stand-alone application have been downloaded from the FTP site for users to run calculations on their local servers; we therefore do not have an accurate count of how many InterProScan searches are run globally per month but can estimate that it must number in the millions. Similarly, the InterPro web site averages around 8 million hits a month from over 50 000 unique hosts.

Despite the high usage statistics that we see, we also recognize the importance of utilizing the latest trends and technologies to make data more readily available to our users. Our intention is to redesign our website to make it more navigable to the novice user and allow more complex querying of the data by advanced users. To help us in our design decisions, a user survey has been carried out to identify features that users like or dislike and to discover what is missing from the resource; the results of the survey will drive future database development. We will provide more data via our web interface, including visualization of UniParc matches and we intend to release our protein match data on a more frequent basis, in synchronization with UniProtKB. As well as improving our web interface, we also aim to increase the amount of data available to users via SOAP and REST-based web services, thus reducing the need for data to be provided in static flat files on the FTP site. We aim to continue to give InterPro's data a functional, structural and evolutional context to ensure its continued usefulness to the biological community.

FUNDING

European Union (213037); Biotechnology and Biological Sciences Research Council (BB/F010508/1); National Institute of Health (GM081084); Wellcome Trust (to AB., R.D.F. and J.M.). Funding for open access charge: European Bioinformatics Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,B., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al*. (2007) New developments in the InterPro database. *Nucleic Acids Res*., **35**, D224–D228.
2. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J.A. (2006) The PROSITE database. *Nucleic Acids Res*., **34**, D227–D230.
3. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al*. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*., **31**, 400–402.
4. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,J.S., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res*., **36**, D281–D288.
5. Bru,C., Courcelle,E., Carrère,S., Beausse,Y., Dalmar,S. and Kahn,D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res*., **33**, D212–D215.
6. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*., **34**, D257–D260.
7. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res*., **31**, 371–373.
8. Nikolskaya,A.N., Arighi,C.N., Huang,H., Barker,W.C. and Wu,C.H. (2006) PIRSF family classification system for protein functional and evolutionary analysis. *Evol. Bioinform. Online*, **2**, 197–209.
9. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res*., **35**, D308–D313.
10. Mi,H., Guo,N., Kejariwal,A. and Thomas,P.D. (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res*., **35**, D247–D252.

11. Yeats,C., Lees,J., Reid,A., Kellam,P., Martin,N., Liu,X. and Orengo,C. (2008) Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res.*, **36**, D414–D418.
12. UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
13. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
14. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2004) UniProt Archive. *Bioinformatics*, **20**, 3236–3237.
15. Brayton,K.A., Lau,A.O.T., Herndon,D.R., Hannick,L., Kappmeyer,L.S., Berens,S.J., Bidwell,S.L., Brown,W.C., Crabtree,J., Fadrosh,D. *et al.* (2007) Genome Sequence of Babesia bovis and Comparative Analysis of Apicomplexan Hemoprotozoa. *PLoS Pathogens*, **3**, e148.
16. Itoh,T., Tanaka1,T., Barrero,R.A., Yamasaki,C., Fujii,Y., Hilton,P.B., Antonio,B.A., Aono,H., Apweiler,R., Bruskiewich,R. *et al.* (2007) Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. *Genome Res.*, **17**, 175–183.
17. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
18. Rawlings,N.D., Tolle,D.P. and Barrett,A.J. (2004) MEROPS: The peptidase database. *Nucleic Acids Res.*, **32**, D160–D164.
19. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct – Open Source Resource for Molecular Interaction Data. *Nucleic Acids Res.*, **35**, D561–D565.
20. Petryszak,R., Kretschmann,E., Wieser,D. and Apweiler,R. (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609.
21. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acid Res.*, **35**, D301–D303.
22. Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32**, D217–D222.
23. Kopp,J. and Schwede,T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res.*, **34**, D315–D318.
24. Labarga,A., Valentin,F., Anderson,M. and Lopez,R. (2007) Web Services at the European Bioinformatics Institute. *Nucleic Acids Res.*, **35**, W6–W11.
25. Prlic,A., Down,T. and Hubbard,T.J.P. (2005) Adding some SPICE to DAS. *Bioinformatics*, **21**, ii40–ii41.
26. Jiminez,R.C., Quinn,A.F., Garcia,A., Labarga,A., O'Neill,K., Martinez,F., Salazar,G.A. and Hermjakob,H. (2008) Dasty2, an ajax protein DAS client. *Bioinformatics*, **24**, 2119–2121.